

Summer Institute for Statistics of Big Data

University of Washington

PIs: Ali Shojaie and Daniela Witten

Grant Number: 1R25EB020380-01

Funding is sought for the Summer Institute for Statistics of Big Data (SISBID) at the University of Washington. This program will provide workshops on the statistical and computational skills needed to access, process, manage, and analyze large biomedical data sets. It will be co-directed by Ali Shojaie and Daniela Witten, faculty in the Department of Biostatistics at University of Washington. The SISBID program will consist of five 2.5-day in-person courses, or modules, taught at the University of Washington each July. An individual participant can register for whichever set of modules he or she chooses. The five modules are as follows: (1) Accessing Biomedical Big Data; (2) Data Visualization; (3) Supervised Methods for Statistical Machine Learning; (4) Unsupervised Methods for Statistical Machine Learning; (5) Reproducible Research for Biomedical Big Data. Each module will consist of a combination of formal lectures and hands-on computing labs. Participants will work together in teams in order to apply the skills that they develop in each module to important problems drawn from relevant case studies. The primary audience for SISBID will consist of biomedical scientists who would like to develop the statistical and computational training needed to make use of Biomedical Big Data. The secondary audience will consist of individuals with stronger statistical or computational backgrounds but little exposure to biology, who will learn how to apply their skills to problems associated with Biomedical Big Data. Participants will include advanced undergraduates, graduate students, post-doctoral fellows, and researchers, and will be drawn from industry, government, and academia. In order to ensure that all participants are able to fully engage in the program, participants will be expected to already have some prior background in R programming and statistical inference, which can be obtained by taking two free online courses before the program begins. Each of the five modules will be co-taught by two instructors. The ten instructors will be drawn from top universities and research centers across the U.S., such as the University of Washington, Rice University, University of Iowa, Johns Hopkins University, MD Anderson Cancer Research Center, Fred Hutchinson Cancer Research Center, and University of North Carolina. They have been selected based on research expertise and excellence in teaching. Lecture videos and slides will be made freely available online so that individuals who are unable to attend SISBID in person can still benefit from the program. This proposal specifically requests funds for 55 student / postdoctoral fellow travel scholarships per year, 130 student / postdoctoral fellow registration scholarships per year, instructor travel and stipends, teaching assistant stipends, and PI salary support. PUBLIC HEALTH RELEVANCE: In recent years, the biomedical sciences have been inundated by Big Data, such as DNA sequence data and electronic medical records. In principle, it should be possible to use such data for a variety of tasks, such as predicting an individual's risk of developing diabetes or cancer, and tailoring therapies to an individual should he or she become ill. The Summer Institute for Statistics of Big Data will provide biomedical researchers with the computational and statistical training needed in order to take advantage of Big Data, so that they can more effectively use it to understand human diseases and to improve human health.